Learning Games

著者	HANAKI Nobuyuki, ISHIKAWA Ryuichiro, AKIYAMA Eizo
year	2007-12
シリーズ	Department of Social Systems and Management Discussion Paper Series ~ no. 1187
URL	http://hdl.handle.net/2241/100201

Department of Social Systems and Management

Discussion Paper Series

No.1187

Learning Games

by

Nobuyuki HANAKI, Ryuichiro ISHIKAWA, and Eizo AKIYAMA

December 2007

UNIVERSITY OF TSUKUBA

Tsukuba, Ibaraki 305-8573

JAPAN

Learning Games*

Nobuyuki Hanaki[a]†, Ryuichiro Ishikawa[b], Eizo Akiyama[b]‡

[a] *Graduate School of Humanities and Social Sciences,*
[b] *Graduate School of Systems and Information Engineering,*
University of Tsukuba, 1-1-1 Ten-nodai, Tsukuba 305-8573, Japan

December 1, 2007

Abstract

This paper proposes a model of learning about a game. Players initially have little knowledge about the game. Through playing the identical game repeatedly, each player not only learns which action to choose but also constructs his personal view on the game. The model is studied using a hybrid payoff matrix of the prisoner's dilemma and coordination games. Results of computer simulations show (1) when all the players are slow in learning the game, they have only a partial understanding of the game, but may enjoy higher payoffs than the cases with full or no understanding of the game; (2) when one of the players is quick in learning the game, he obtains a higher payoff than the others. However, all of them can receive lower payoffs than the case where all the players are slow learners.

Keywords: Learning, Subjective views, Computer simulation

JEL code: C72, D83

*We thank Mamoru Kaneko, Jeff Kline, and seminar participants at GREQAM for comments and suggestions. This research is partially supported by the Japan Society for the Promotion of Sciences, Grant-in-Aid for Scientific Research (S), No. 17103002, and by the Ministry of Education, Culture, Sports, Science and Technology, Grants-in-Aid for Young Scientists (B), No. 19730137 and No. 19730164.

†Corresponding Author. Tel: +81-29-853-7432. Fax: +81-29-853-7440.

‡E-mail addresses: hanaki@dpipe.tsukuba.ac.jp (N. Hanaki), ishikawa@sk.tsukuba.ac.jp (R. Ishikawa), and eizo@sk.tsukuba.ac.jp (E. Akiyama).

1 Introduction

In standard game theory, players are assumed to have well-formed beliefs and knowledge of the structure of the game they play. The origin of their beliefs and knowledge is rarely studied. The validity of this assumption is, however, questioned in experimental economics.[1] According to Camerer (2003, p. 474), "what game do people think they are playing?" is one of the top 10 most important open research question in experimental game theory.

If players do not understand a game completely, or misunderstand it, how do they learn about the true game? In the extensive literature, both theoretical and experimental, on learning in games, this question has been seldom addressed. The literature has mainly focused on learning about how to play a game rather than on learning about the game itself.[2] An exception is Oechssler and Schipper (2003). They conducted a set of experiments in which subjects did not know the payoffs of their opponent and were given incentive to learn about them in 2×2 games. The authors constructed the games that subjects perceived they were playing—the subjective games—from the data. They found that the subjective games differed frequently from the games actually being played.

In this paper, we study a model in which players play a normal form game repeatedly and learn not only about how to play the game but also about the game itself. A normal form game consists of the set of players, the set of available actions (or strategies), and the payoff function for each player. Therefore, learning about a game means that players do not know some of these components and learn about them. In this paper, we assume a player knows about the set of actions available to himself and the number of opponents. However, initially he does not know about the set of actions available to his opponents or anyone's payoff functions. The player learns about them—in particular, his own payoffs associated with possible outcomes—by playing the game repeatedly. In this paper, therefore, players

[1] Camerer (2003, p. 474) gives an example of a student who participated in an experiment at Caltech. The students confused the coordination game used in the experiment with the prisoner's dilemma game, and "defected" continuously.

[2] Fudenberg and Levine (1998) provided a detailed survey of the theoretical literature of learning in games. See, for example, Crawford (1995), Cheung and Friedman (1997), Mookherjee and Sopher (1997), Erev and Roth (1998), and Camerer and Ho (1999) for the experimental learning literature. Hanaki, Sethi, Erev, and Peterhansl (2005) proposed a model in which players learn about which repeated game strategies to use.

are learning about different aspects of a game than those that were studied in Oechssler and Schipper (2003).

To model how players build their personal views about a game from playing it repeatedly, we have adopted an idea from cognitive science, namely, the role that *autobiographical memory* plays in learning from everyday life events (Linton, 1982; Wagenaar, 1986). An autobiographical memory is a memory of frequently repeated events. It is an abstract script so that the details, such as the date of occurrence, are lost, and only the general facts about the events remain. In order to replicate such a way of learning about events in our model, players are assumed to have two types of memories, *short-term and long-term memories*.[3]

A short-term memory is a temporary memory of an outcome of playing the game, i.e., the actions chosen by players and the payoff received by the player.[4] The short-term memory remains in the player's mind only for a certain number of periods, and vanishes after that. If the same outcome is repeated frequently enough while short-term memories of it remain in the player's mind, the outcome will be kept as a long-term memory. In other words, an outcome of the game will stay in the player's mind as a long-term memory, if it has been experienced frequently enough during the specified periods. A long-term memory permanently associates, in the mind of the player, an outcome of the game with a payoff. Once an outcome of the game is engraved in the player's mind as a long-term memory, it remains there forever, and we say that he has learned the part of the game corresponding to it. The personal view of a player about the game is simply the part of the game he has learned.

In addition to learning about the payoffs, players learn which action to choose. The latter is modeled based on the reinforcement learning model. When a player does not know any of the payoffs, only the realized payoffs will be utilized. As the player learns some parts of the game, he starts to infer what the payoffs could have been if he had

[3]The model of the mind's memory system composed of short-term and long-term memories was first proposed by James (1890) and established by Atkinson and Shiffrin (1968). Long-term memory can be classified into episodic memory and semantic memory (Tulving, 1972). Autobiographical memory is a type of episodic memory for information related to *oneself* (Brewer, 1986).

[4]In general, the word "memory" can mean either a storage of information or information itself. In this paper, we mainly use "memory" in the latter sense.

acted differently at least for the parts of the game he knows. Therefore, learning about performance of actions will be based not only on the realized payoffs but also on the forgone payoffs where possible. We have studied this model in a 3×3 game that embeds both a prisoner's dilemma and a coordination game. Through a series of computer simulations of the model, we demonstrate what kind of personal views the players tend to form and what kind of behavior emerges when players' personal views and their behavior coevolve.

We find that, when all the players are very slow in learning about a game, they will only have a limited understanding (a partial view) of the game. Such a limited understanding, however, can be beneficial for them. They may enjoy higher payoffs than the case of full or no understanding of the game. Because the players enjoy high payoffs, their behaviors do not change. Because their behaviors do not change, neither do their personal views. Therefore, their views remain partial. When one of the players is quick in learning about a game, he can obtain a higher payoff than the other players who are slow in learning. However, in this case, all the players, even the fast-learning player, may obtain lower payoffs than in the case where all the players are very slow learners.

The rest of the paper is organized as follows. Section 2 presents the model in detail. Results of the model simulation are summarized in Section 3. Section 4 offers an explanation of the results, and Section 5 concludes.

2 A Model of Learning about a Game

We consider a two-person game. The set of players is $\{1, 2\}$, and each player $i \in \{1, 2\}$ has the set S^i of available actions. Initially, each player only knows the set of actions that is available to him, but he does not know the action sets of the others nor the payoffs, i.e., he does not know the game he is facing. Through playing the game repeatedly, players not only learn which action will bring about higher payoffs but also form their views of the game they are playing. Below, we first discuss how we model the formation of personal views by players, and the representation of such views. Once we define a player's personal view of the game, we formulate how the player learns which action to choose based on the past outcomes and his view of the game.

2.1 Formation of personal views

We assume that a player has two types of memories, *short-term and long-term memories*. A short-term memory is a memory of an outcome of playing the game, i.e., the actions chosen by players and the payoff received by the players. The short-term memory remains in the player's mind only for a certain number of periods, and vanishes after that. If the same outcome is repeated frequently enough while short-term memories of it remain in the player's mind, it will be kept as a long-term memory. Once an outcome of the game is engraved in the player's mind as a long-term memory, it remains there forever, and we say that he has learned the part of the game corresponding to it.

More precisely, player i is characterized by his short-term memory length m^i and cognition threshold $k^i (\leq m^i)$. The short-term memory length is the number of periods before a short-term memory vanishes from his mind. The cognition threshold represents the number of repetitions needed for an outcome to be kept as a long-term memory. Because a short-term memory vanishes after m^i periods, an outcome (s^i, s^j) will stay in i's mind as a long-term memory, if it has been experienced k^i times in the m^i most recent interactions. Once the outcome (s^i, s^j) is recorded as a long-term memory, the player knows the payoff he can receive if the outcome is realized in the future.

The transformation of an outcome of the game in the mind of a player from a short-term memory to a long-term memory in this paper plays a similar role as in the *autobiographical memory* in cognitive science. The autobiographical memory is the memory of everyday life events. As shown by Linton (1982) and Wagenaar (1986), this memory keeps a repeated event as an abstract script. That is, when keeping such repeated events, details such as date of occurrence are lost, and only the general facts about the events remain. In addition, if an event is not repeated, the event will not remain as an autobiographical memory.

In this paper, a player's personal view of the game will be defined based both on the objective payoff matrix and the set of long-term memories in his mind. Because the set of long-term memories in the player's mind may change over time, so does his personal view of the game.

Let Π be the objective payoff matrix of the game under consideration, and let Π^i represent the part of the payoff matrix that corresponds to what player i receives. Namely, in a two person game,

$$\Pi^i = \begin{pmatrix} \pi^i(s_1^i, s_1^j) & \dots & \pi^i(s_1^i, s_{n^j}^j) \\ \vdots & \ddots & \vdots \\ \pi^i(s_{n^i}^i, s_1^j) & \dots & \pi^i(s_{n^i}^i, s_{n^j}^j) \end{pmatrix},$$

where n^i and n^j are the numbers of actions in S^i and S^j, respectively, and $\pi^i : S^i \times S^j \to \mathbf{R}$ is player i's payoff function. We assume that $\pi^i(s^i, s^j) \neq 0$ for all $s^i \in S^i, s^j \in S^j, i, j \in \{1, 2\}$ with $j \neq i$. This is because we assign a special meaning to value zero in the subjective payoff matrix as defined below.

Let $L^i(t)$ be the matrix that represents the state of the long-term memories in the mind of player i at period t, where each element of the matrix takes value zero or one; $L^i_{s^i,s^j}(t) \in \{0, 1\}$. $L^i_{s^i,s^j}(t)$ takes value zero when outcome $\left(s^i, s^j\right)$ is not in player i's mind as a long-term memory at period t, and it is one otherwise. We assume that initially players do not know about any of the outcomes; that is, $L^i_{s^i,s^j}(0) = 0$ for all $(s^i, s^j) \in S^i \times S^j$.

The personal view of the game for player i at period t, $\widetilde{\Pi}^i(t)$, can be defined as:

$$\widetilde{\Pi}^i(t) = L^i(t) \cdot \Pi^i. \tag{1}$$

Therefore, $\widetilde{\Pi}^i_{s^i,s^j}(t)$ is zero when player i has not learned of the outcome $\left(s^i, s^j\right)$ at period t, and it is equal to $\pi^i(s^i, s^j)$ otherwise. We call this matrix the subjective payoff matrix for player i at period t. Now we proceed to discuss how players learn which action to choose.

2.2 Learning about performance of actions

We assume that a player's recent experiences from choosing (as well as not choosing) an action are summarized by his "attraction" for the action. In each period, players choose their actions based on their attractions for each action. It is through the evolution of attractions that players learn.

Let $A_s^i(t)$ denote player i's attraction for action $s \in S^i$ at period t. The probability that player i chooses action s at period t, $p_s^i(t)$, depends on the player's attraction as follows:

$$p_s^i(t) = \frac{e^{\lambda^i A_s^i(t)}}{\sum_{k \in S^i} e^{\lambda^i A_k^i(t)}}. \tag{2}$$

The parameter λ^i in the logistic transformation represents the extent to which actions with higher attractions are favored in action choice. When $\lambda^i = 0$, all actions are equally likely to be chosen regardless of their attraction. As λ^i becomes larger, actions with higher attractions become disproportionally more likely to be chosen. In the limiting case where $\lambda^i \to \infty$, the action with the highest attraction is chosen with probability one. The logistic transformation introduced here is common in the literature on learning in games as well as experimental game theory to model better the action choices of subjects in laboratory experiments (see, for example, McKelvey and Palfrey, 1995; Erev and Roth, 1998; Camerer, 2003). We assume that all the actions have the same attraction for all the players at the beginning of the game, i.e., $A_s^i(0) = 0$ for all i and $s \in S^i$. Therefore, initially, all the actions are equally likely to be chosen regardless of λ^i.

Attractions evolve as follows:

$$A_s^i(t+1) = \frac{1}{h^i} \sum_{\tau=0}^{h^i-1} R_s^i(t-\tau), \tag{3}$$

where $h^i = \min(m^i, t+1)$.[5] $R_s^i(t)$ is a stimulus the player receives for action s at period t, which depends on the outcome of the game as well as the player's understanding of the game in period t in the following manner:

$$R_s^i(t) = \begin{cases} \pi^i(s^i(t), s^j(t)) & \text{if } s = s^i(t) \\ \widetilde{\Pi}^i_{s,s^j(t)}(t) & \text{otherwise,} \end{cases} \tag{4}$$

where $s^i(t)$ represents the action chosen by player i in period t. Equation (4) states that the stimulus player i receives for action $s \in S^i$ at t is the realized payoff when he chooses s at period t, i.e., $s = s^i(t)$, regardless of the status of long-term memories in i's mind. If

[5] $h^i = \min(m^i, t+1)$ is to take care of the early periods so that the game has not been played m^i times.

the player i does not choose s at period t, then the stimulus follows i's subjective payoff matrix. Therefore, if the payoff consequence of $(s, s^j(t))$ is in i's mind as a long-term memory, then the stimulus for action s will be the forgone payoff, i.e., the payoff player i could have obtained if he had chosen action s in period t, given the action chosen by the opponent in that period, $s^j(t)$. Otherwise, the stimulus will be zero.

In this definition of stimulus, we are assuming that once a player understands some of the payoffs of the game, he can infer what the payoffs could have been if he had acted differently for the part of the game he knows. For the parts of the game he does not know, he cannot make such inferences.

The proposed model of learning about performance of actions builds on two models of action learning commonly studied in the literature: learning based only on realized payoffs (see, e.g., Erev and Roth, 1998) and learning based on both forgone and realized payoffs (see, e.g., Camerer and Ho, 1999). The bridge between these two in our model is the long-term memories or personal views. Indeed, when $L^i_{s^i,s^j}(t) = 0$ for all t, $(s^i, s^j) \in S^i \times S^j$, players in our model learn about performance of actions based only on realized payoffs. In contrast, when $L^i_{s^i,s^j}(t) = 1$ for all t, $(s^i, s^j) \in S^i \times S^j$, players always learn based on both realized and forgone payoffs.[6] It is the dynamics of long-term memories that makes our model different from existing learning models. In the next section, we show the results of computational simulation of our model.

3 Simulation Results

We consider symmetric 3×3 games in this paper. Each player $i \in \{1, 2\}$ has three available actions $\{s^i_1, s^i_2, s^i_3\}$. The objective payoff matrix Π with a parameter $a \in (0, 0.5)$ is given as follows.

[6] As shown by Camerer and Ho (1999), when the learning is based both on realized and forgone payoffs, the behavior the model generates will be equivalent to the one generated by fictitious play with probabilistic action choice.

	s_1^2	s_2^2	s_3^2
s_1^1	$1-a$, $1-a$	0, 1	1, 0
s_2^1	1, 0	a, a	a, 0
s_3^1	0, 1	0, a	$1-a$, $1-a$

As mentioned in the previous section, we have given a special meaning to the zero in the subjective payoff matrix, the payoff consequences of outcomes that are not kept as long-term memories. In order not to have zero in the objective payoff matrix, we added $b = 0.01$ to all the payoffs.[7] Note that b is not shown in the above payoff matrix for clarity of exposition. The unique pure Nash equilibrium of this game is (s_2^1, s_2^2) with payoff (a, a).

The game is constructed so that a prisoner's dilemma game (four cells in the upper left corner) and a coordination game (four cells in the lower right corner) are embedded. This is to investigate the effect of personal views in our model. As players learn the game, they may see themselves facing a prisoners' dilemma-type situation, a coordination game-type situation, or something else. Depending on how players understand the game, their behaviors may vary. The parameter a determines the severity of the dilemma in the prisoner's dilemma game as well as the risk–payoff trade-off in the coordination game. In the prisoner's dilemma game embedded here, the lower a is, the larger the aggregate loss of not choosing (s_1^1, s_1^2) becomes. In the embedded coordination game, if $1/3 < a < 0.5$, (s_2^1, s_2^2) is the risk-dominant equilibrium while (s_3^1, s_3^2) is the payoff-dominant equilibrium. It is interesting to see what kind of views players construct and what kind of behavior they learn over time under various values of a.

In the simulation analysis below, we first focus on the cases where all the players have the same short-term memory length, $m^i = m$, and the same cognition threshold, $k^i = k$. We then proceed to the cases where players have the same short-term memory length but different cognition thresholds. Throughout the paper, we assume that the sensitivity of action choices to the attractions are the same across players, $\lambda^i = \lambda$ for all $i \in \{1, 2\}$.

[7] The results remain the same if we subtract $b = 0.01$ from all the payoffs.

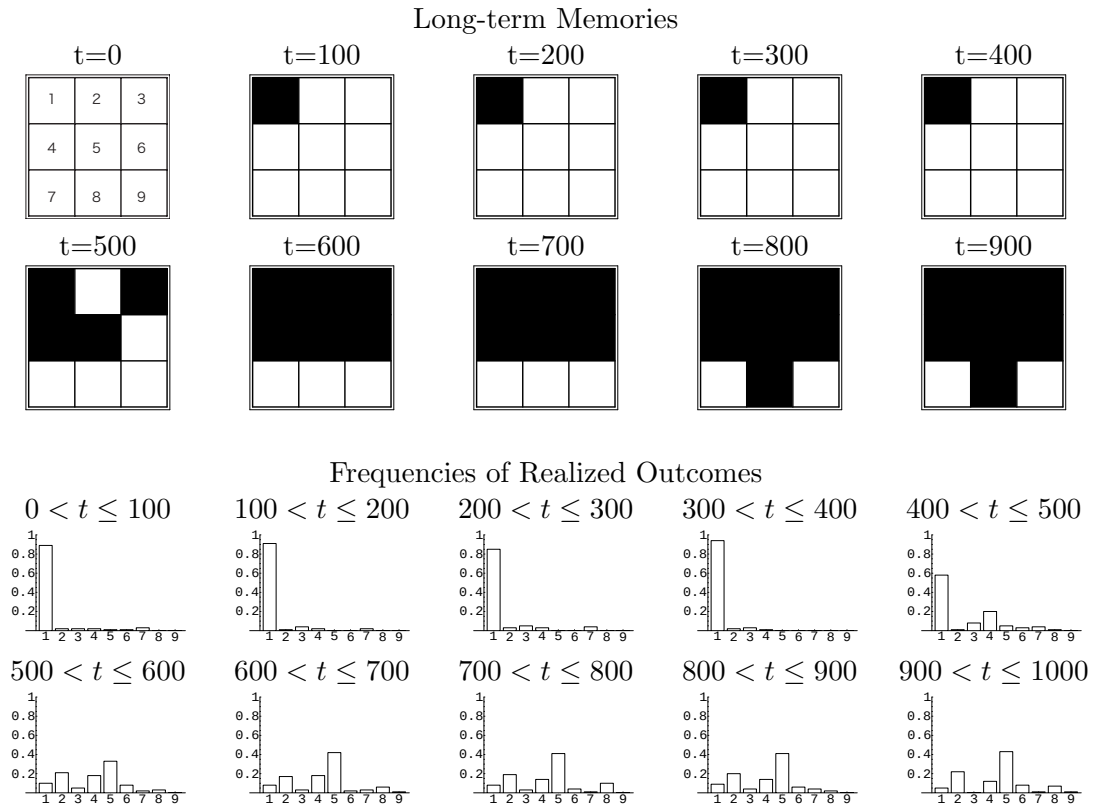

Figure 1: The evolution of the long-term memory (top) and the corresponding frequencies of realized outcomes (bottom) from a single simulation run. $m = 5$, $k = 3$, $\lambda = 5.0$, $a = 0.25$. The black cells represent those outcomes recorded as long-term memories.

3.1 Case of identical memory length and threshold

We first consider the case where the short-term memory length of the players and their cognition thresholds are identical across them. In such cases, we have four parameters in our model: a defines the payoff matrix, m and k determine the length of the players' short-term memory and their cognition threshold, and λ governs the importance of attractions in action choices. We will show results based on a particular set of parameter values ($m = 5$, $\lambda = 5.0$, $a = 0.25$ while varying k). The dependencies of the results on the parameter values are discussed in the appendix.

Figure 1 shows an outcome of a simulation run.[8] The parameters are set so that $m = 5$, $k = 3$, $\lambda = 5.0$, and $a = 0.25$. To illustrate the evolution of players' long-term memories

[8] A single simulation run consists of 1000 interactions by a pair of players. One period in the simulation corresponds to one interaction by the pair of players.

(top panel), the outcomes that are kept as long-term memories at a given point in time are shown by the black cells. Nine cells in the objective payoff matrix are numbered from 1 to 9 as shown in the status of long-term memories at $t = 0$. The frequencies of realization of each outcome of the game are represented by the height of the corresponding bars in the bottom panel. In this simulation run, outcome 1 has been realized with a high frequency in earlier periods. As a result, it became a long-term memory of the two players. Between period 400 and 500, however, because of a player's deviation from playing action 1, the players learned other outcomes. Consequently, the behaviors of the players change quite drastically in the later periods. As one can see from the figure, in the later period, outcome 5, which is the Nash equilibrium of the game, is realized with the highest frequency.

The result in Figure 1 is just an example of how the players' understanding of the game and their behaviors coevolve as players repeatedly play the game. However, it is not the representative result. In fact, there can be many other patterns of coevolution. Instead of enumerating all the possible results, we focus on averaged results[9] below.

Figure 2 shows the results of simulations for $a = 0.25$. In each figure, the average payoffs of the row-player over time (top),[10] the average status of long-term memories at $t = 500$ (bottom), and the average frequencies of realized outcomes for $500 \leq t \leq 1000$ (middle) are shown for $k = 1$ (left), $k = 3$ (center), and $k = 5$ (right). Except for the average status of the long-term memories, the outcome of the two models of learning, the one based only on realized payoff (RL model) and the other based on both realized and forgone payoff (FP model) are reported.[11]

In the top panel of the figure, the outcomes from the three models are shown. The average payoff of the row-player from the RL model ("RL players") is shown with solid gray, and that from the FP model ("FP players") is shown with dashed gray. The result of our model (game learning model, "GL players") is shown in solid black. The middle panel shows, by the height of the bars, the frequencies with which outcomes corresponding

[9]For each set of parameter values, we take the average of the results generated by 100 simulation runs, while giving varying random seeds for each run.

[10]The average payoff of the column-player is very similar to that of the row-player.

[11]The two models considered in this paper are those with a probabilistic action choice and fixed memory length. In both of these models, actions are chosen based on the attraction following equation (2) while the attraction evolves following equation (3) except that the status of long-term memories is modified as discussed in the main text above.

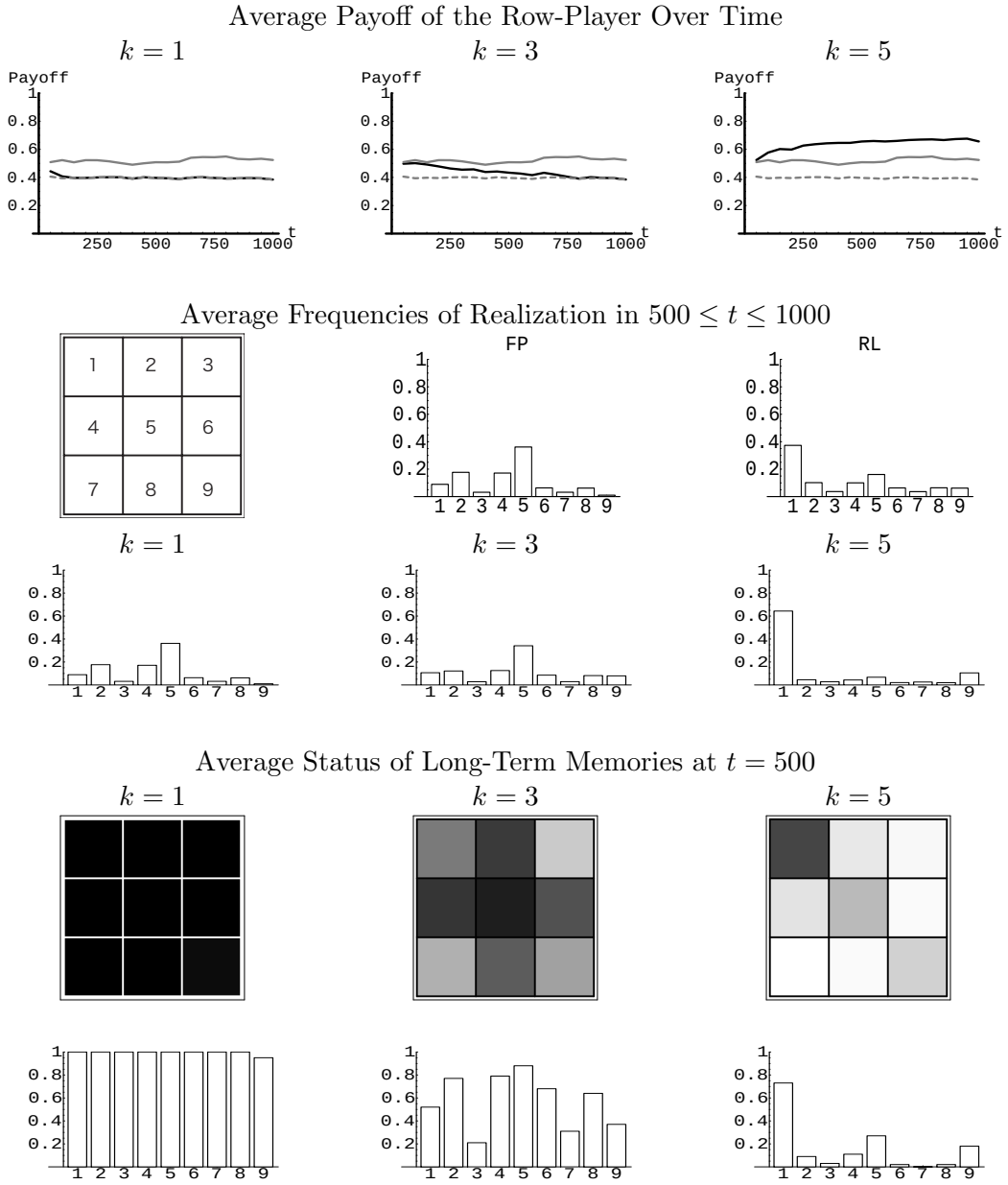

Figure 2: Average payoff of the row-player over time (top), average frequencies of realized outcomes (middle), and average status of long-term memories (bottom) for various k. $m = 5$, $a = 0.25$, $\lambda = 5.0$. For the average payoff, the result of our model is in solid black, the solid gray represents the RL model, and the dashed gray represents the FP model. For the average status in the long-term memories, the darker gray corresponds to the higher likelihood that the outcome is kept as a long-term memory, which is also shown by the height of bars.

to cells numbered from 1 to 9 are realized. The FP in the figure stands for the outcome of the "FP model" while RL stands for that of the "RL model". In the bottom panel, the darkness of a cell and the corresponding height of the bars show the proportion of the simulation runs for which the outcome was recorded as a long-term memory at $t = 500$.

When $k = 1$, one can see that the payoff received by GL players quickly converges to those received by FP players. When $k = 1$, players learn all the payoffs of the game quite quickly. As one can see in the figure, by $t = 500$, all the payoffs are known by the players almost all the time.[12] Once GL players learn of all the payoffs, their behaviors become equivalent to those of FP players.

The convergence in the average payoff is much slower in the case of $k = 3$. In this case, players do not learn all the payoffs as in the $k = 1$ case. Players learn, however, the outcomes 2, 4, 5, 6, and 8 by $t = 500$ in the majority of the simulation runs. The result shows that such a partial understanding is enough for players to choose the Nash equilibrium action with a high probability.

Do partial understandings of the game always lead players to choose the Nash equilibrium action with a higher likelihood? The results from the $k = 5$ case shows that the answer is no. When $k = 5$, GL players receive higher payoffs than both RL and FP players. The partial understandings of the game that resulted in this case gives more benefits to players than the full or no understanding of the game. The figure shows that players learn outcome 1, which Pareto dominates Nash equilibrium outcome 5, in the majority of simulation runs. While outcomes 5 and 9 are also learned, such cases are infrequent.

The result comes from both a partial view of the game and the players' limited ability to acquire the view. As mentioned in Section 2, it is hard for GL players with high cognition thresholds, k, to keep the outcomes as a long-term memory. Outcome 1 realized by these players is stable once it becomes a long-term memory. This is because it is hard for player 1 (player 2) to encounter outcome 4 (2) repeatedly enough to learn that it is more attractive than outcome 1, which is already engraved in his mind as a long-term memory.

It should be noted that players' action choices and their understandings of the game

[12] In fact, when $k = 1$, players understand all the payoffs by period 100 in most of the simulations.

coevolve in our model. Therefore, not only do players benefit from their limited understanding of the game, but also, because they benefit from such a limited understanding, their behaviors do not change. Therefore, their views remain partial.[13]

This result is quite interesting and illuminates the possibility that as we live in a very complex society, it may not be feasible for us to learn the true or complete interactive environment that we face. Our understanding of the environment may be very limited, but as long as we are satisfied with the outcomes, we do not actively try to learn the true environment (or do not try and see what will happen if we do something different from what we normally do). Therefore, our understandings remain limited. Of course, it is quite possible that, because of our limited understanding and the lack of exploration, we are not receiving a higher payoff, which could be obtainable if we really understand the complete environment.

3.2 Case of identical memory length but different thresholds

In the previous subsection, we have seen the case where players have identical short-term memory length m and cognition threshold k. When all the players are slow in constructing their personal view of the game (k close to m), they can obtain a higher payoff than when they are quick in learning the game structure (small k). What happens if two players who have the same short-term memory length $m^1 = m^2 = m$, but different cognition thresholds, interact? We consider such cases in this subsection.

Figure 3 shows the typical dynamics of long-term memory of two players and the frequencies of realized outcomes when player 1's cognitive threshold is 1, $k^1 = 1$, and that of player 2 is 5, $k^2 = 5$. Both the players have the same short-term memory length, $m = 5$. The sensitivity of action choices to attractions, λ, are set equal to 5, and the payoff matrix is such that $a = 0.25$.

In this particular simulation run, both players learn outcomes 1 and 5 by period 100. While player 2 only learns these two, player 1 learns all the outcomes except for outcome 9. By period 500, player 1 learns all the outcomes, but player 2's understanding still remains

[13]The dependency of the results on parameter values are discussed in the appendix, which shows that the result holds in quite a large parameter space as long as players' cognition thresholds, k, are high enough, i.e., close to their short-term memory lengths, m.

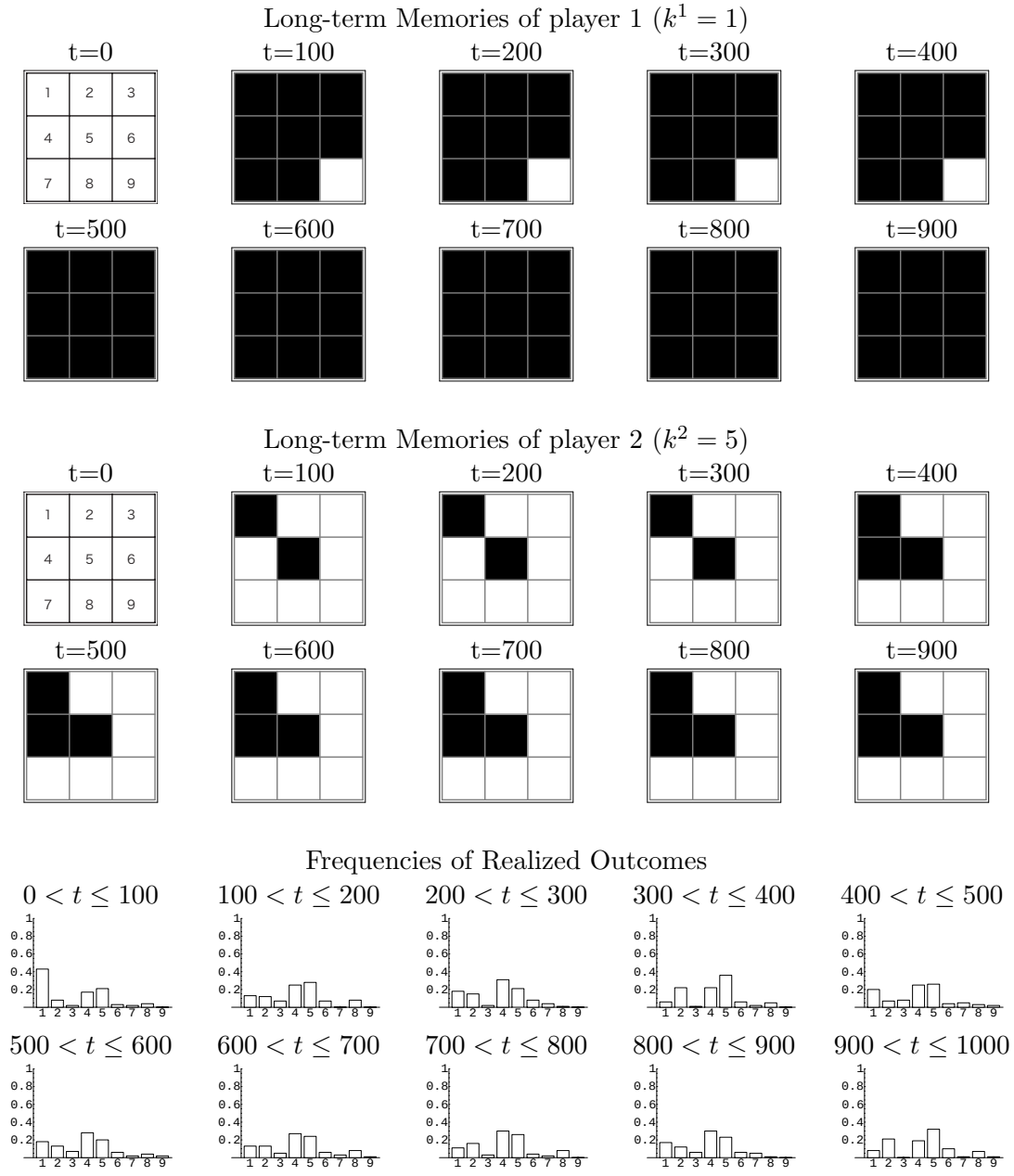

Figure 3: The evolution of the long-term memory for a player with low k, $k^1 = 1$, (top) and high k, $k^2 = 5$ (middle) and the corresponding frequencies of realized outcomes (bottom) from a single simulation run. $m = 5$, $\lambda = 5.0$, $a = 0.25$. The black cells represent the outcomes that are kept as long-term memories.

limited to outcomes 1, 4, and 5.

The dynamics of the frequencies of the realized outcomes are quite interesting. In the first 100 periods, it was outcome 1 that had been realized the most. Beyond period 100, outcomes 4 and 5 are realized with higher frequencies than outcome 1. It should be noted that because player 1 knows almost all the payoffs, when both players are happily choosing action 1 (therefore outcome 1 is realized), player 1 can infer that if he chooses action 2, while player 2 keeps choosing action 1, he can get a higher payoff (associated with outcome 4). Therefore, there is a high chance that he will change his behavior. However, if player 1 indeed chooses action 2, player 2 may think that it is better to choose action 2 instead of action 1. (Recall that player 2 knows the payoff associated with outcome 5!) Such learning will result in both players indeed choosing action 2, i.e., the Nash equilibrium.

This example shows that even if only one of the players is quick in learning about the game, two players may learn to choose the actions that correspond to the Nash equilibrium outcome. It is not necessary that both the players be quick in learning about the game to reach the Nash equilibrium.

Figure 4 shows the averaged results[14] of the simulation runs. It shows, for three different thresholds ($k = 1, 3, 5$) of player 1, average payoffs of players over time (top), average frequencies of realized outcomes (middle), and average status of long-term memories for two players (bottom). In these simulation runs, player 2's thresholds are fixed at $k^2 = 5$. Other parameter values are $m^1 = m^2 = 5$, $a = 0.25$, and $\lambda = 5.0$. We are presenting the $k^1 = k^2 = 5$ case as the benchmark.

In the top panel, the average payoff of the player with a lower k, player 1, is the solid line, while that of the player with a high k, player 2, is the dashed line. In the figure, player 1 receives a higher payoff than player 2. This is because a player who is quick in learning about the game, and therefore better understands the game, can make more sophisticated decisions than the other player.

However, interestingly, compared with the case where both players have a very high cognitive threshold, $k^1 = k^2 = 5$, payoffs for *both players* are lower when one of the players

[14] As noted above, for each set of parameter values, we take the average of the results generated by 100 simulation runs, while giving varying random seeds for each run.

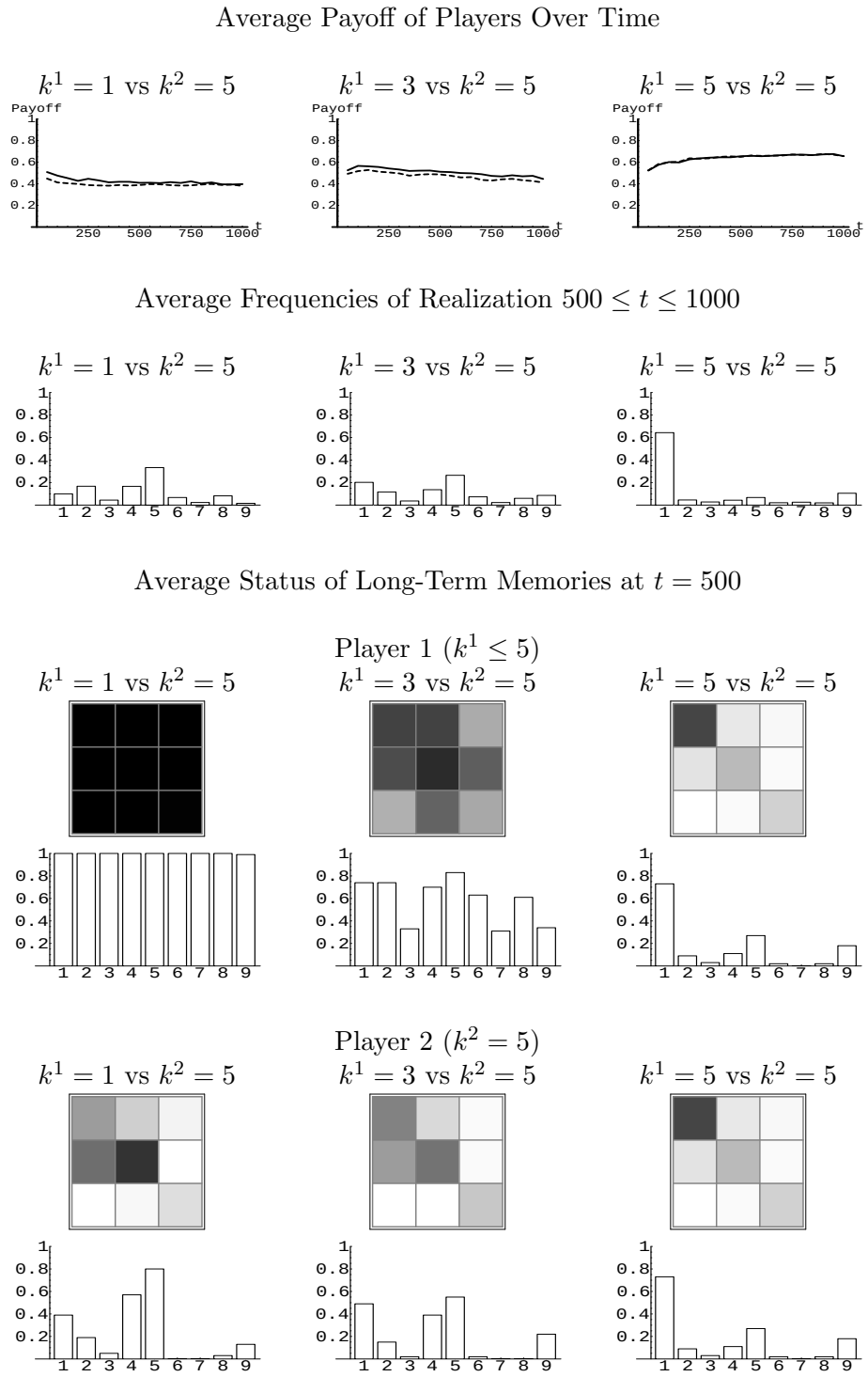

Figure 4: Average payoff of the row-player over time (top), average frequencies of realized outcomes (middle), and average status of long-term memories (bottom) for three values of k^1. $m^1 = m^2 = 5$, $k^2 = 5$, $a = 0.25$, $\lambda = 5.0$. In the top figure, the average payoff of player 1 (low cognitive threshold) is the solid line, while that of player 2 (high cognitive threshold) is the dashed line. For the average status in the long-term memories, the darker gray corresponds to the higher likelihood that the outcome is kept as the long-term memory. The same information is also shown by the height of bars.

has a lower cognitive threshold. As seen above, the player with a low cognitive threshold, player 1, learns almost all the outcomes while the player with a high threshold, player 2, learns outcome 1. As player 1 takes action 2, the payoff that player 2 receives from using action 1 decreases. As a result, player 2 also learns to take action 2 to get a higher payoff. Once both the players learn the Nash equilibrium outcome, they do not deviate from it. However, because the Nash equilibrium outcome is Pareto dominated, the payoffs of both the players are lower.

Note also that when one of the players has a low k, the player with a high k also learns the Nash equilibrium outcome much more often (see the bottom panel of the figure). In addition, when a is smaller, the payoff difference between the player with a low k and the player with a high k is larger. See the appendix for more discussion.

4 An Account for the Simulation Results

Why can a partial understanding of the game structure benefit players in the focal game, and why do behaviors of players remain such that their understanding of the game remains partial? In this section, we provide an explanation through a highly simplified analysis of the case where players have an identical short-term memory length and cognition threshold.

In our simulation, the model always seems to reach a steady state where probabilities with which players choose their actions do not vary over time (at least, if we take averages over several realizations). Here we restrict our analysis to such an average steady state. Because the game in this paper is symmetric and the players have identical characteristics (i.e., m^i, k^i, and λ^i are all the same across players), we focus on a symmetric steady state. More rigorous and exact analysis is needed to understand fully the behavior of the model, but the simplified analysis below can explain the main result in our simulation analysis, namely, why partial understanding of a game can benefit players and why behaviors of players remain such that their understandings of the game remain partial.

Let us start with considering the learning based only on the realized payoffs (RL

model). The expected steady-state level of attractions, A_s^i, for action s in this case is:

$$A_s^i = p_s^i \left(p_1^j \pi(a_s^i, a_1^j) + p_2^j \pi(a_s^i, a_2^j) + p_3^j \pi(a_s^i, a_3^j) \right), \tag{5}$$

where p_s^i and p_s^j are player i's and j's probabilities for choosing action s, respectively, when $t \to \infty$ in equation (2).

This is because, in this model, an action will not receive the stimulus unless it is actually chosen, and when it is chosen, the expected value of the stimulus that the action receives depends on the probabilities with which the opponent is choosing the actions. Note that $p_s^j = \frac{\exp(\lambda A_s^j)}{\sum_{k=1}^3 \exp(\lambda A_k^j)}$. Because, in the symmetric steady state, $A_s^i = A_s^j (\equiv A_s^{RL})$ for $s \in \{1, 2, 3\}$, the expected levels of steady-state attractions for the RL model (A_s^{RL}) become:

$$\begin{aligned} A_1^{RL} &= p_1^{RL}(p_1^{RL}(1-a) + p_3^{RL}), \\ A_2^{RL} &= p_2^{RL}(p_1^{RL} + p_2^{RL}a + p_3^{RL}a), \\ A_3^{RL} &= p_3^{RL}(p_1^{RL} + p_3^{RL}(1-a)), \end{aligned}$$

where $p_s^{RL} = \frac{\exp(\lambda A_s^{RL})}{\sum_{k=1}^3 \exp(\lambda A_k^{RL})}$ for each action s in the game. Here we have ignored b added to the payoffs for clarity of exposition. Solving these equations for A_1^{RL}, A_2^{RL}, and A_3^{RL} gives us the expected levels of attractions for each action in the steady state, and the expected probabilities that players choose each action in the steady state follows immediately.[15]

In our model of game learning, players' understanding of the game plays a role in determining the expected steady-state attractions. For example, if the payoff associated with (s_1^i, s_1^j) is the unique long-term memory in players' minds, then the steady-state attractions of our model (A_s^{GL}) become:

$$\begin{aligned} A_1^{GL} &= p_1^{GL}(1-a) + p_1^{GL}p_3^{GL}, \\ A_2^{GL} &= p_2^{GL}(p_1^{GL} + p_2^{GL}a + p_3^{GL}a), \\ A_3^{GL} &= p_3^{GL}(p_1^{GL} + p_3^{GL}(1-a)). \end{aligned}$$

[15] It is possible that there are multiple steady states.

This is because action 1 will receive stimulus not only when outcomes (s_1^i, s_1^j), (s_1^i, s_2^j), and (s_1^i, s_3^j) are realized but also every time the opponent chooses action 1. One can expect from these equations that when players only learn of the payoffs associated with (s_1^i, s_1^j), action 1 will have a higher steady-state attraction than the RL model and will be chosen with a higher probability by players. As a result, (s_1^i, s_1^j) will be observed much more frequently in our model than in the case of the RL model, and because players choose other actions with a low probability, other outcomes are not realized frequently enough. Therefore, their understanding of the game remains partial.

On the other hand, if players quickly learn the entire game, as in the $k = 1$ case, the steady-state attractions of our model immediately become equivalent to those of learning based on both realized and forgone payoffs (FP model). The expected attraction for action s in the FP model is given by:

$$A_s^i = p_1^j \pi(a_s^i, a_1^j) + p_2^j \pi(a_s^i, a_2^j) + p_3^j \pi(a_s^i, a_3^j). \tag{6}$$

This is because all the actions will always receive stimulus regardless of whether they have been chosen or not. Therefore, the expected levels of steady-state attractions in the FP model (A_s^{FP}) become:

$$\begin{aligned} A_1^{FP} &= p_1^{FP}(1-a) + p_3^{FP}, \\ A_2^{FP} &= p_1^{FP} + p_2^{FP} a + p_3^{FP} a, \\ A_3^{FP} &= p_1^{FP} + p_3^{FP}(1-a), \end{aligned}$$

for the game when we ignore the b added to the payoffs. These expressions give the largest weight to choosing action 2 for both of the players. Therefore, compared with the RL model, the FP model results in players obtaining lower payoffs.

5 Summary and Conclusion

In this paper, we have presented a model of learning about a game. Players initially have little knowledge about the game they play. They gain experience through playing the

game repeatedly, and based on their experience, they not only learn which action will bring about a higher payoff but also form their view about the game they are playing. We show that, in the 3×3 game we have considered, which embeds both a prisoner's dilemma and a coordination game, players may benefit from having a very limited understanding of the game when all the players have such a limited understanding. Their payoffs can be higher than the cases where players have full or no understanding of the game. It should be noted that personal views and behaviors of players coevolve in our model. Because players enjoy a high payoff, their understanding of the game remains partial and vice versa. This result suggests that players may live happily without fully understanding highly complex strategic environments.

When one of the players has a much better understanding of the game than the other, the one with the better understanding can enjoy a higher payoff than the one with less understanding. However, their payoffs—even the payoffs of the player who better understands the game—can be lower than in the case where all the players have a limited understanding. The behavior of the player who better understands the game can lead the other player to respond in a way that lowers their payoffs. This, combined with the results above, suggests that a benefit of ignorance may exist, but it exists only when everyone is ignorant.

In this paper, we have considered a pair of players playing the game repeatedly. However, one can easily extend the framework presented here to the case where there are many players to be matched with a few others. In such a case, it is possible to consider various matching protocols: for example, players may be situated in a network and interact only locally. The results here suggest that it is possible that players may form several different "local views" of the same objective game. What will happen when there is occasional random matching among those with different views? Are there views that can spread much more easily than others? These are all interesting questions to investigate, but we will leave them for future research.

It is also interesting to conduct laboratory experiments and examine how subjects learn in the situation considered in this paper; namely, subjects are initially only informed about the set of actions available to themselves, but they observe actions chosen by all

the relevant players and the payoff received after each interaction. Do subjects behave in the way that the model predicts? We also leave these questions for future research.

References

Atkinson, R., and R. Shiffrin (1968): "Human memory: A proposed system and its control processes," in *The psychology of learning and motivation: Advances in research and theory*, ed. by K. Spence, and J. Spence, vol. 2, pp. 89–195. Academic Press.

Brewer, W. F. (1986): "What is autobiographical memory?," in *Autobiographical memory*, ed. by D. Rubin, pp. 25–49. Cambridge University Press.

Camerer, C., and T.-H. Ho (1999): "Experience-weighted attraction learning in normal form games," *Econometrica*, 67, 827–874.

Camerer, C. F. (2003): *Behavioral Game Theory: Experiments in Strategic Interaction*. Russell Sage Foundation, New York, NY.

Cheung, Y.-W., and D. Friedman (1997): "Individual learning in normal form games: Some laboratory results," *Games and Economic Behavior*, 19, 46–76.

Crawford, V. P. (1995): "Adaptive dynamics in coordination games," *Econometrica*, 63, 103–143.

Erev, I., and A. E. Roth (1998): "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, 88, 848–881.

Fudenberg, D., and D. K. Levine (1998): *The Theory of Learning in Games*. MIT Press, Cambridge, MA.

Hanaki, N., R. Sethi, I. Erev, and A. Peterhansl (2005): "Learning strategy," *Journal of Economic Behavior and Organization*, 56, 523–542.

James, W. (1890): *The principles of psychology*. Henry Holt and Co.

Linton, H. A. (1982): "Transformations of memory in everyday life," in *Memory observed: Remembering in natural contexts*, ed. by U. Neisser. Freeman.

McKelvey, R. D., and T. R. Palfrey (1995): "Quantal response equilibria for normal form games," *Games and Economic Behavior*, 10, 6–38.

Mookherjee, D., and B. Sopher (1997): "Learning and decision costs in experimental constant sum games," *Games and Economic Behavior*, 19, 97–132.

Oechssler, J., and B. Schipper (2003): "Can you guess the game you are playing?," *Games and Economic Behavior*, 43, 137–152.

Tulving, E. (1972): "Episodic and semantic memory," in *Organization of memory*, ed. by E. Tulving, and W. Donaldson, pp. 381–403. Academic Press, New York.

Wagenaar, W. A. (1986): "My memory: A study of autobiographical memory over six years," *Cognitive Psychology*, 18, 225–252.

A Dependency of the Results on Parameter Values

In the main text, we presented results under $m^1 = m^2 = 5$, $a = 0.25$, $\lambda = 5.0$. Here we show what happens to the results if we change the parameter values.

Figures 5 and 6 show the results for the case where $a = 0.05$ and $a = 0.45$, respectively, in the same format as Figure 2. As in the case discussed in the main text, when $k = 1$, GL players quickly learn all the outcomes, and their behaviors converge to that of FP players. When $a = 0.45$, there is not much difference in behavior among the GL, FP, and RL players. It is also the case that among GL players, the differences in cognition threshold k do not affect their behavior significantly. In all the models, players learn to play the Nash equilibrium, and they learn the payoff associated with the Nash equilibrium outcome almost all the time.

In these two figures, the GL players do not receive higher payoffs than FP or RL players, contrary to what we have shown in the main text and Figure 2. What is the range of a over which our main result holds? How about the range of λ? Figure 7 shows the average payoff of players for various values of a holding λ constant at $\lambda = 5.0$ (top) as well as various values of λ while holding a constant at $a = 0.25$. One can see that the GL players receive higher payoffs than FP and RL players over quite a large parameter space, in particular, $0.1 \leq a \leq 0.35$ when $\lambda = 5.0$ and $3.0 \leq \lambda \leq 6.0$ for $a = 0.25$. In fact, we have experimented with other values of m, and qualitatively the same results can be obtained: as k becomes closer to m, the GL players receive high payoffs while their understanding of the game remains very limited, although the specific values of a and λ for which such a result holds depend on m.

We have seen that when players have the same short-term memory length but different cognition thresholds, the player with a low cognition threshold (the one who learns the game quickly) receives a higher payoff than the one with a high cognition threshold. The difference between the payoffs received by the two players is larger when the difference between the two thresholds is large and also when a is low.

Figure 8 shows the results for simulation runs when $m^1 = m^2 = 5$, $\lambda = 5.0$ and $a = 0.05$. As one can see, the lower is k^1, the larger is the payoff difference between

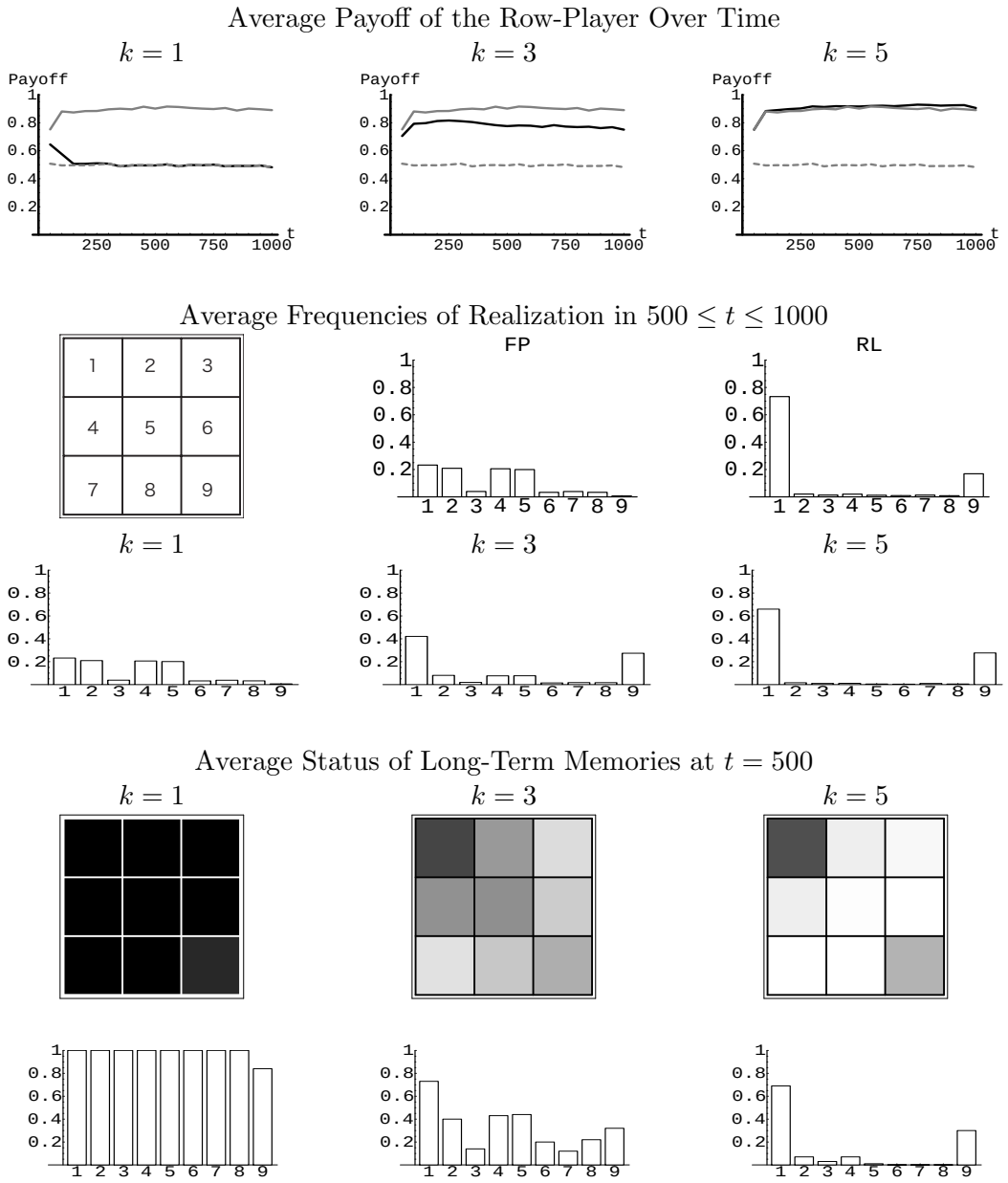

Figure 5: Average payoff of the row-player over time (top), average frequencies of realized outcomes (middle), and average status of long-term memories (bottom) for various k. $m = 5$, $a = 0.05$, $\lambda = 5.0$. For the average payoff, the result of our model is in solid black, the solid gray represents the RL model, and the dashed gray represents the FP model. For the average status in the long-term memory, the darker gray corresponds to the higher likelihood that the outcome is kept as a long-term memory, which is also shown by the height of bars.

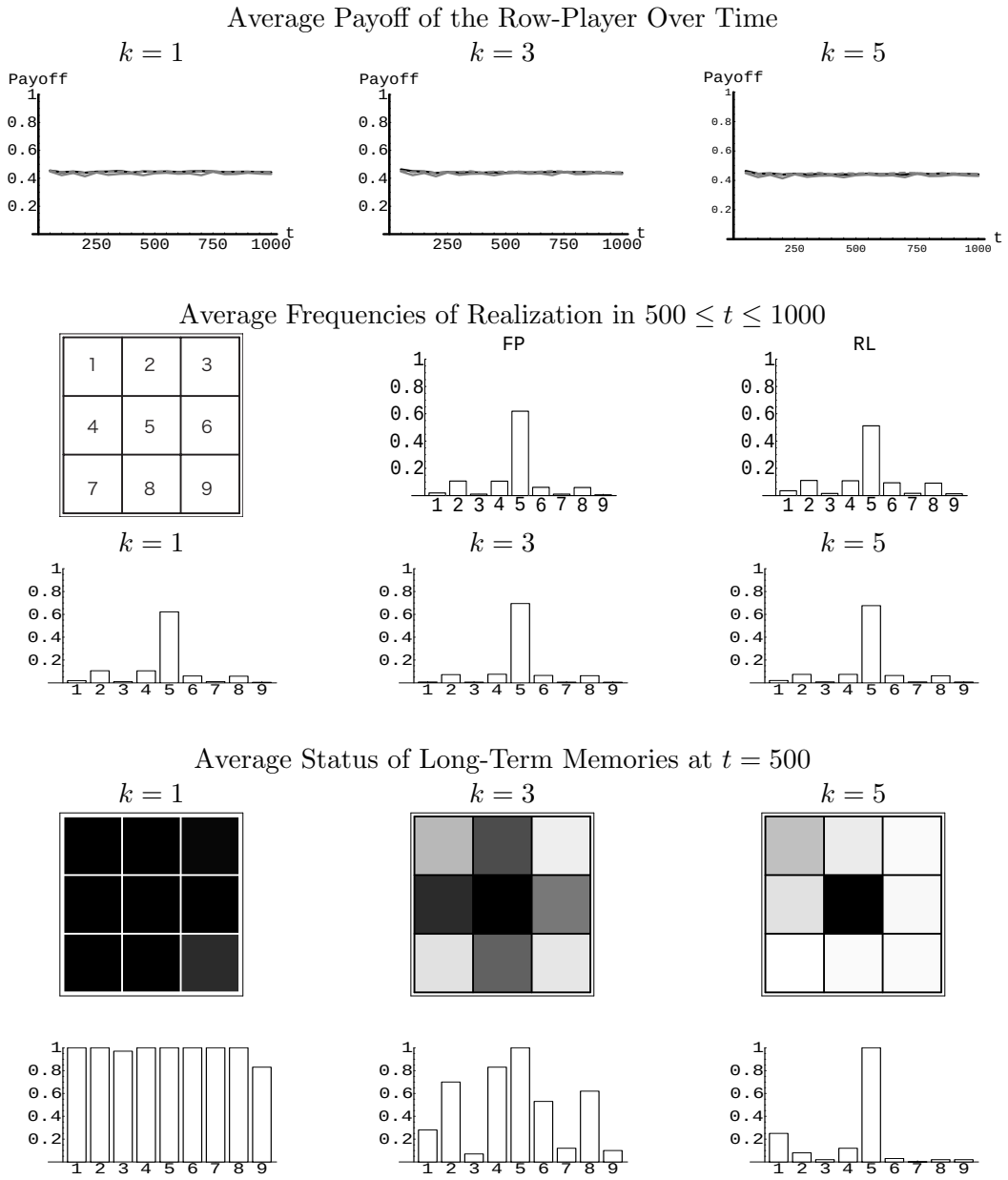

Figure 6: Average payoff of the row-player over time (top), average frequencies of realized outcomes (middle), and average status of long-term memories (bottom) for various k. $m = 5$, $a = 0.45$, $\lambda = 5.0$. For the average payoff, the result of our model is in solid black, the solid gray represents the RL model, and the dashed gray represents the FP model. For the average status in the long-term memory, the darker gray corresponds to the higher likelihood that the outcome is kept as a long-term memory. The same information is also shown by the height of bars.

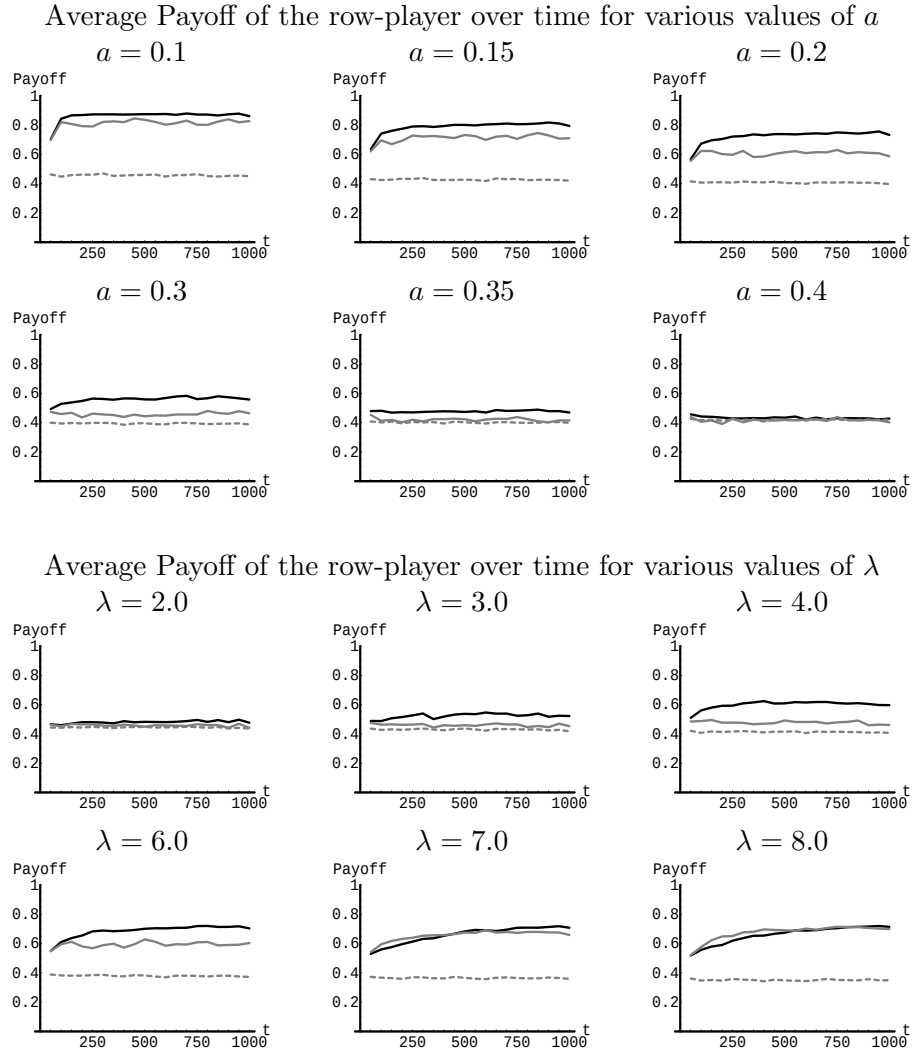

Figure 7: Average payoff of the row-player over time for various a (top) and for various values of λ (bottom). $m = 5$, $k = 5$, $\lambda = 5.0$ (top) and $a = 0.25$ (bottom). For the average payoff, the result of our model is in solid black, the solid gray represents RL model, and the dashed gray represents FP model.

the two players. Furthermore, compared with what was shown in Figure 4, the payoff differences between the two players are larger here.

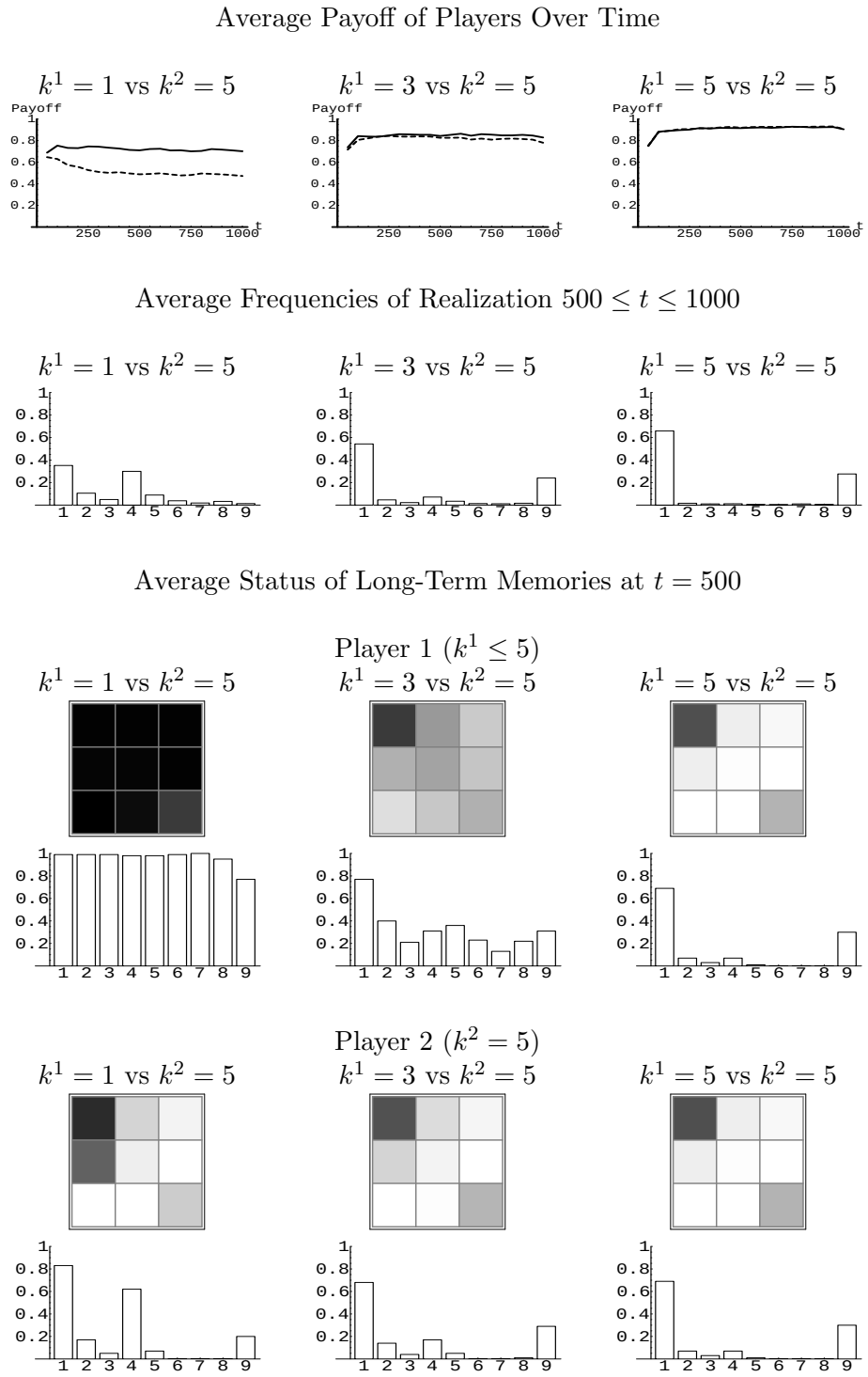

Figure 8: Average payoff of the row-player over time (top), average frequencies of realized outcomes (middle), and average status of long-term memories (bottom) for three values of k^1. $m^1 = m^2 = 5$, $k^2 = 5$, $a = 0.05$, $\lambda = 5.0$. In the top panel, the average payoff of player 1 (low cognitive threshold) is in solid line, while that of player 2 (high cognitive threshold) is in the dashed line. For the average status of the long-term memories, the darker gray corresponds to the higher likelihood that the outcome is recorded as a long-term memory. The same information is also shown by the height of bars.